BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e58982
doi: 10.3897/biss.4.58982

OPEN ACCESS

Conference Abstract

Task Group 2 - Data Quality Tests and Assertions

Lee Belbin‡, Arthur Chapman§, John Wieczorek|, Paul J. Morris¶, Paula F Zermoglio|

‡ Atlas of Living Australia, CSIRO, Canberra, Australia
§ Australian Biodiversity Information Services, Ballan, Australia
| VertNet, Bariloche, Argentina
¶ Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States of America

Corresponding author: Lee Belbin (leebelbin@gmail.com)

Received: 25 Sep 2020 | Published: 01 Oct 2020

Citation: Belbin L, Chapman A, Wieczorek J, Morris PJ, Zermoglio PF (2020) Task Group 2 – Data Quality Tests and Assertions. Biodiversity Information Science and Standards 4: e58982. https://doi.org/10.3897/biss.4.58982

Abstract

Motivation

Other than data availability, 'Data Quality' is probably the most significant issue for users of biodiversity data and this is especially so for the research community. Data Quality Tests and Assertions Task Group (TG-2) from the Biodiversity Information Standards (TDWG) Biodiversity Quality Interest Group is reviewing practical aspects relating to 'data quality' with a goal of providing a current best practice at the key interface between data users and data providers: tests and assertions. If an internationally agreed standard suite of core tests and resulting assertions can be used by all data providers and aggregators and hopefully data collectors, then greater and more appropriate use could be made of biodiversity data. Adopting this suite of core tests, data providers and particularly aggregators such as the Global Biodiversity Information Facility (GBIF) and its nodes would have increased credibility with the user communities and could provide more effective information for evaluating 'fitness for use'.

Goals, Outputs and Outcomes

- A standard core (fundamental) set of tests and associated assertions based around Darwin Core terms
- A standard suite of descriptive fields for each test
- Broad deployment of the tests, from collector to aggregator
- A set of basic principles for the creation of tests/assertions

© Belbin L et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

- Software that provides an example implementation of each test
- Data that can be used to validate an implementation of the tests
- A publication that captures the knowledge built during the creation of the tests/ assertions

Strategy

The tests and rules generating assertions at the record-level are more fundamental than the tools or workflows that will be based on them. The priority is to create a fully documented suite of core tests that define a framework for ready extension across terms and domains.

Status 2019-2020

The core tests have proven to be far more complex than any of the team had anticipated. Several times over the past three years, we believed we had finalized the tests, only to find new issues that have required a fresh understanding and subsequent edits, e.g., the most recent dropping of the two tests related to dwc:identificationQualifier:

- TG2-VALIDATION_IDENTIFICATIONQUALIFIER_DETECTED and
- TG2-AMENDMENT_IDENTIFICATIONQUALIFIER_FROM_TAXON

This decision resulted from a review of dwc:identificationQualifier values in GBIF records and an evaluation of expected values based on the Darwin Core definition of the term. Aside from there being many values, the term expects the qualifier in relation to a given taxonomic name, and rules of open nomenclature are unevenly adopted across data records to reliably parse and detect dwc:identificationQualifier for these tests to be effective.

A similar situation occurs for dwc:scientificName, where we have resorted to the term “polynomial” to refer to the non-authorship part of dwc:scientificName.

What has occurred during the past year?

- Months of work on discussions and edits to the GitHub issues (= mainly the tests), using mainly via Zoom and email.
- We had hoped to have a face-to-face meeting in Bariloche, Argentina early in 2020 but the Corona virus stopped that. This was unfortunate as we needed this meeting to discuss the remaining complex issues as noted above. Attempting to address such issues by Zoom has been far less efficient.
- We are occasionally re-visiting decisions made years earlier. An indication that we have been doing this work for (too) many years.
- We have now standardized all the test parameters for the 99 CORE tests. Much work has gone into standardizing the phrasing and terminology within the 'Expected response' field of the tests – the parameter that most clearly defines each test.
- Two of the test fields that have taken most of our time to resolve have been ‘Parameters’ and what we now call ‘bdq:sourceAuthority’ (Chapman et al. 2020a).

These are now complete. The work on 'Parameters' has fed in to Task Group 4 on Vocabularies of Values (see Vocabularies needed for Darwin Core terms prepared by TG4).

- We have published the work from the Data Quality Interest and Task Groups: Chapman et al. 2020b
- We have extended the vocabulary that has been used for the Tests and Assertions.
- Development of the datasets that validate the implementation of the tests continues.
- We recognize the dependence on the work of the Annotations Interest Group for the results from the tests to have maximal impact. It is important that test results stay with the records.

We will provide details of the challenges, the breakdown of the tests and the advances of the project.

Keywords

Darwin Core, fitness-for-use, parameters, code

Presenting author

Lee Belbin

Presented at

TDWG 2020

References

- Chapman A, Wieczorek J, Belbin L, Morris P, Veiga AK (2020a) Vocabulary of terms used for the TDWG Task Group on Data Quality Tests and Assertions. 10.3897/biss.4.50889.suppl1
- Chapman A, Belbin L, Zermoglio P, Wieczorek J, Morris P, Nicholls M, Rees ER, Veiga A, Thompson A, Saraiva A, James S, Gendreau C, Benson A, Schigel D (2020b) Developing Standards for Improved Data Quality and for Selecting Fit for Use Biodiversity Data. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.50889